Task Group	Data Center Bridging
Revision	1.01
Author	Manoj Wadekar (Qlogic), et al

Priority Grouping for DCB Networks

(Enhanced Transmission Selection)

Rev 1.01

Modification History

Rev	Originator	Comment
1.0	Manoj Wadekar	Initial Submitted Version
1.01	Manoj Wadekar	Added clarification text for PGID=15 behavior Added clarification on granularity of BW check

1 Authors

The following people, with company affiliations, have contributed to the preparation of this proposal:

Amit Shukla – Juniper
Anoop Ghanwani - Brocade
Anjan – Cisco
Anthony Faustini - Cisco
Asif Hazarika – Fujitsu
Avi Godbole – Juniper
Awais Nemat – Marvell
Bruce Klemin – Qlogic
Brice Kwan - Broadcom
Claudio DeSanti- Cisco
Craig W. Carlson - QLogic
Dan Eisenhauer – IBM
Danny J. Mitzel - Brocade
David Peterson – Brocade
Diego Crupniokoff – Mellanox
Dinesh Dutt - Cisco
Douglas Dreyer - IBM
Ed Bugnion - Nuova
Ed McGlaughlin – Qlogic
Eric Multanen - Intel
Gaurav Chawla - Dell
Glenn - Brocade
Hemal Purohit - QLogic
Hugh Barrass – Cisco
Ilango Ganga - Intel
Irv Robinson - Intel
J. R. Rivers – Nuova
Jeelani Syed - Juniper
Jeffrey Lynch - IBM
Jim Larsen - Intel
Joe Pelissier - Cisco
John Hufferd – Brocade
John Terry - Brocade
Manoj Wadekar – Qlogic
Menu Menuchehry - Marvell
Mike Ko – IBM
Mike Krause - HP
Parag Bhide - Emulex
Pat Thaler - Broadcom
Ravi Shenoy - Emulex
Renato Recio - IBM
Robert Snively - Brocade

Roger Hathorn - IBM
Sanjaya Anand – Qlogic
Sanjay Sane – Cisco
Shreyas Shah - PLX
Silvano Gai - Nuova
Stuart Berman - Emulex
Suresh Vobbilisetty - Brocade
Taufik Ma - Emulex
Uri Elzur - Broadcom

2 Overview

This document provides definitions and an operational model for priority processing and bandwidth allocation on converged links in end stations and switches in a DCB (Data Center Bridging) environment. Using priority-based processing and bandwidth allocations, different traffic classes within different traffic types such as LAN, SAN, IPC, and management can be configured to provide bandwidth allocation, low-latency, or best effort transmit characteristics.

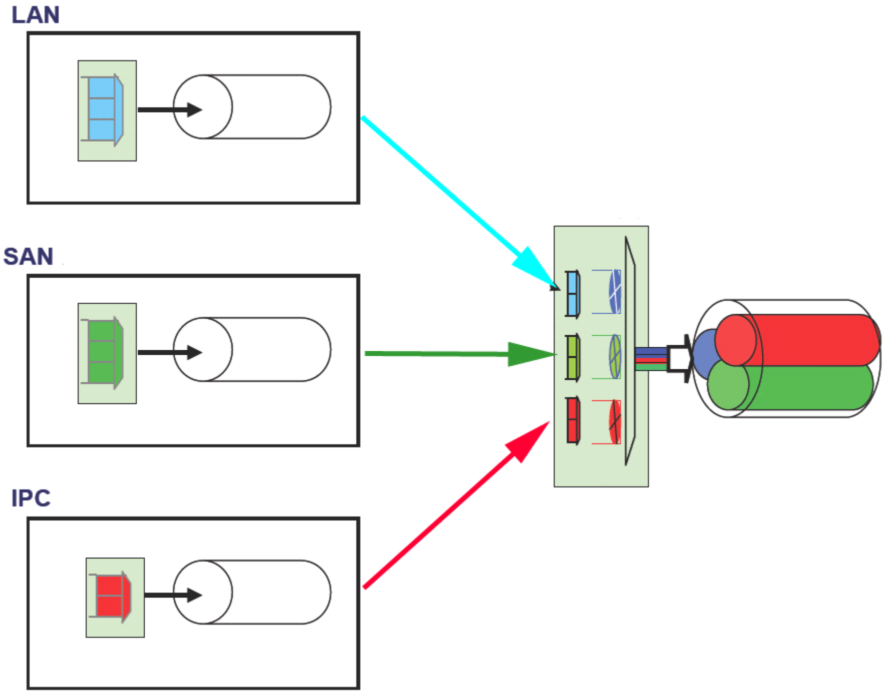

Figure 1: Convergence over Ethernet

3 Goals for enhanced transmission selection in 802.1Q Bridges

With DCB and other new usage models, 802.1Q Bridges needs to service traffic types with requirements different to those on classical 802.1Q Bridges. Converged Ethernet will carry traffic types that are sensitive to different kind of handling. E.g. Storage traffic is sensitive to packet loss, while IPC traffic is latency-sensitive. In a single converged link, all these traffic types need to coexist without imposing serious restrictions on each other's performance. To achieve this, DCB devices need to support following:

- Assignment of priorities to “Priority Groups” such that a priority group represents traffic requiring a certain behavior;
    - e.g. LAN, SAN, IPC;
    - Allow multiple priorities within a “Priority Group”.
- Configuration of BW allocation for each “Priority Group”;
    - The allocation is expressed as a percentage of available link bandwidth;
    - Bandwidth allocation resolution is 1%;
        - e.g. 40% LAN, 40% SAN, 20% IPC
    - Bandwidth allocation within a Priority Group is outside of scope of this document.
- Supports minimum scheduler behavior to minimize the impact of converging the different traffic types on a single link;
    - Allow coexistence of traffic types requiring low latency with traffic types that are bandwidth intensive and loss sensitive;
    - Preserves relative prioritization for some traffic types while allowing bandwidth sharing among others.
- Provide a consistent management framework for configuring the above via MIB objects.

4 Definitions

1. **Priority (Pri):** There are eight traffic priorities, determined by 3-bit priority in the 802.1Q tag.
2. **Priority Group :** A Priority Group is group of priorities bound together by management for the purpose of bandwidth allocation. All priorities in a single group are expected to have similar traffic handling requirements with respect to latency and loss (e.g. congestion managed vs non-congestion managed).
3. **Priority Group ID (PGID):** A 4-bit identifier assigned to a priority group. PGID = 15 is a special value that allows priorities to be configured with “No Bandwidth Limit”. PGID values from 8 to 14 are reserved.
4. **Priority Group BW (PG%):** Percentage of available link bandwidth allocated to a particular PGID.

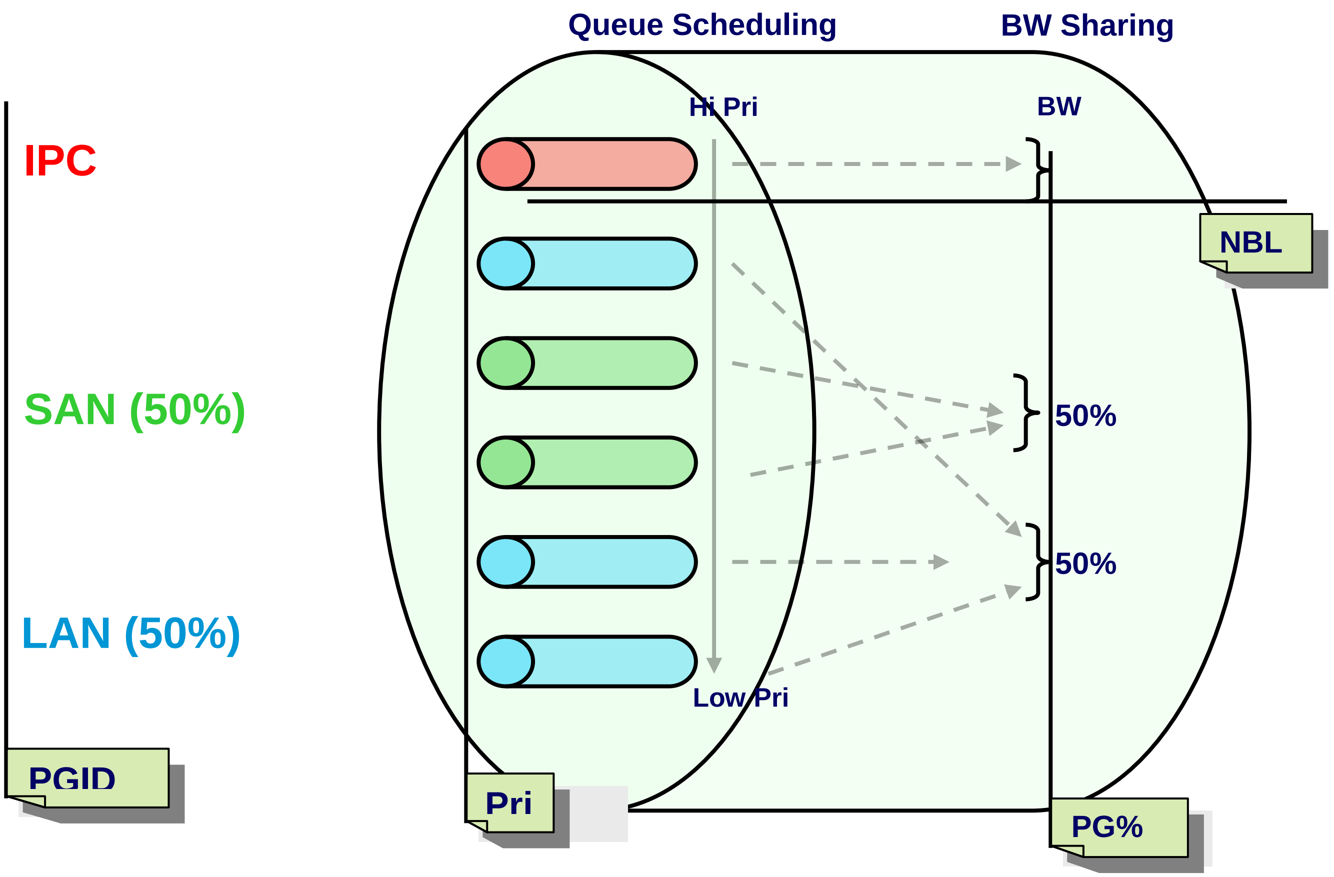

Figure 2: Converged Link Configuration Parameters

5 Configuration Tables

5.1 T1: Pri-PGID Table

Table 1: Mapping of incoming Priority to Priority Groups

Pri	PGID	Desc
0	1	LAN
1	1	LAN
2	0	SAN
3	0	SAN
4	1	LAN
5	1	LAN
6	1	LAN
7	15	IPC

This table[1] binds incoming priority (identified by 802.1Q tag) to PGID within the system. More than one priority value may be mapped to a PGID.

[1] All values in the tables to be considered as an example.

PGID is a 4 bit field with a range of 0-15. A PGID of 15 has a special meaning that priorities mapped to this PGID will not be subjected to BW limit. Any priority that is mapped to PGID 15 is scheduled per the priority/traffic class mechanism as defined currently in 802.1Q-2005 Clause 8.6.8. PGID values 8-14 are reserved.

5.2 T2: PG-BW Table

Table 2: BW Allocation to User Priority Groups

PGID	PG%	DESCRIPTION
15	-	IPC
0	50	SAN
1	50	LAN
-		
-		
-		
-		

This table allows configuration of bandwidth to each Priority Group. The values in the PG% column must sum to 100; behavior is unspecified if this condition is not met.

The Priority Group with PGID 15 must not be allocated any PG%. Priorities belonging to this group are not subjected to BW limit. PGID values 8-14 are reserved. For configuring BW limit, PGID values between 0 and 7 (inclusive) should be used.

To summarize, PGID usage is defined below:
PGID = {0, 7}: To be used when PG is limited for its BW use
PGID = {8, 14}: Reserved
PGID = {15}: To be used for Priorities which should not be limited for their BW use

Configured PG% (PG Percentage in Table 2) refers to the max percentage of available link bandwidth after priorities within PGID 15 are serviced, and assuming that all PGs are fully subscribed. If one of the Priority Groups doesn't consume its allocated bandwidth, then any unused portion is available for use by other Priority Groups.

5.3 Default Legacy Configuration

IEEE 802.1Q specifies strict priority scheduling as default behavior. This can be achieved by configuring each priority in Table 1 to belong to Priority Group with PGID 15. In that case, Table 2 would contain only one Priority Group with no bandwidth allocation.

IEEE 802.1Q specifies Priority to Traffic Class mapping. There is no change to such mapping due to this specification.

6 Compliance Requirements

1. DCB devices shall support at least 3 Priority Groups
    a. DCB devices shall allow configuration of and allocation of one or more priorities to Priority Group 15 (no BW limit).
    b. DCB devices shall allow at least one Priority Group with bandwidth allocation to be configured such that all of the priorities within that group have PFC enabled.
    c. DCB devices shall allow at least one Priority Group with bandwidth allocation to be configured such that all of the priorities within that group have PFC disabled.
2. DCB Devices shall support BW configuration with at least 1% [2] granularity.
3. DCB devices shall support a work conserving transmission selection policy; i.e. if one of the Priority Groups doesn't consume its allocated bandwidth, then any unused bandwidth is available to other priority Groups.

7 Configuration Recommendations

1. DCB administrators should group priorities having similar traffic handling in the same Priority Group; e.g. PFC traffic should not be grouped with non-PFC traffic.
2. DCB administrators should not merge traffic from multiple Priority Groups in same traffic class; doing so may result in undefined behavior.

[2] Specification of precision of scheduler is beyond scope of this document. However we expect precision of scheduler to be between +/- 10%. As a consequence of this, a Priority Group with a PG% of 0% may still receive up to 10% of the available bandwidth, even when the other Priority groups with a non-zero PG% are fully subscribed.